実例を対象とした分類法入門

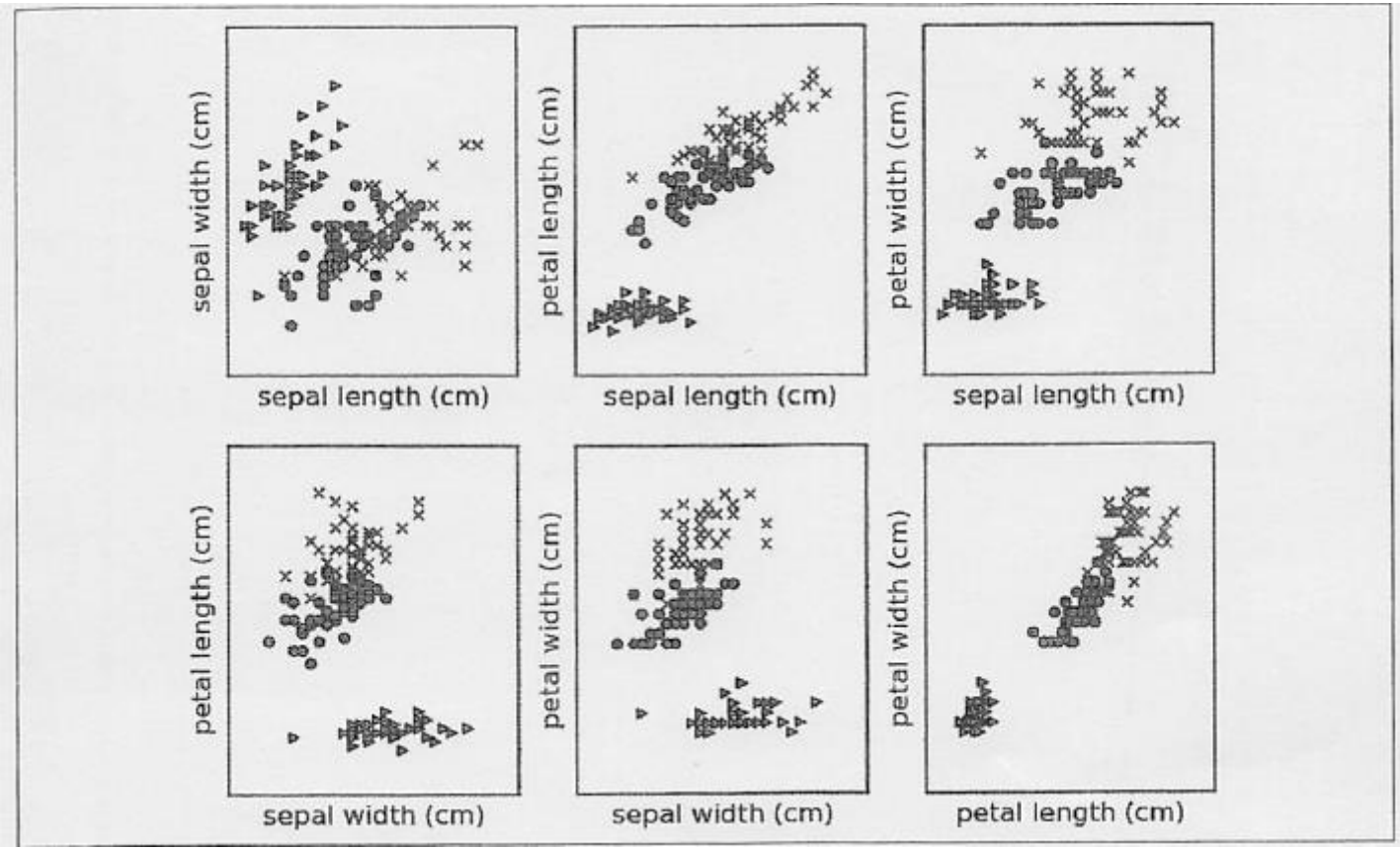
CAO　RUI

アイリスデータセット

- 花の品種:
- Setosa　Versicolor　Virginica

- 計測された要素(特徴量)
  - がく片の長さ
  - がく片の幅
  - 花弁の長さ
  - 花弁の幅

3つ異なる品種の花が二つのグループに分けられ、花弁の長さを用いて、Setosaという品種のアイリスと他の品種を誤りなく見分けることができます



保持データと交差検定で評価を行う

- 先の例

訓練データ・テストデータ

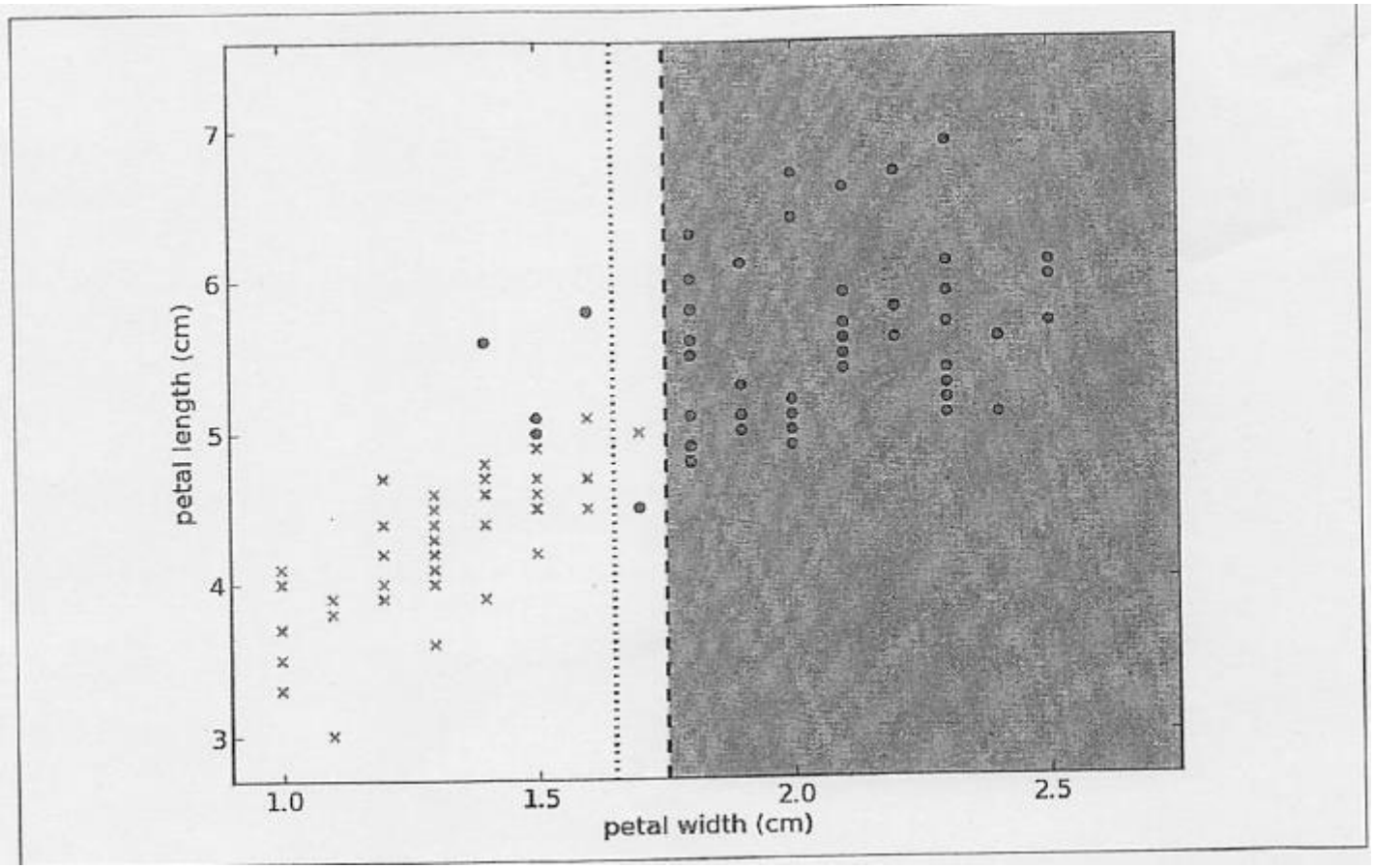
- 正解率：96%
- より信頼性高い結果を求めるため

訓練データ

一部のデータを取り除く

テストデータ

正解率：90%


petal length (cm)
7
6
5
4
3
1.0
1.5
2.0
2.5
petal width (cm)

交差検定とは

- 利点：
- 交差検定の利点は評価を行う時に使ったデータは訓練データと違うものということを保証した上に、訓練データの数がそんなに減らすこともあんまり起こらない
- 欠点：処理時間が増加した
- leave-one-out（処理時間が一番長い）
- K-分割交差検定（一番よい結果）

より複雑なデータセットとクラス分類

- 面積（A）
- 周囲長さ（P）
- 密集度（C=4π A/P2）（特徴量エンジニアリング:システム性能に大きな影響を及ぼす）
- 長さ
- 幅
- 非対称係数
- 殻溝の長さ

小麦の品種:

カナダ産

コマ産

ロサ産

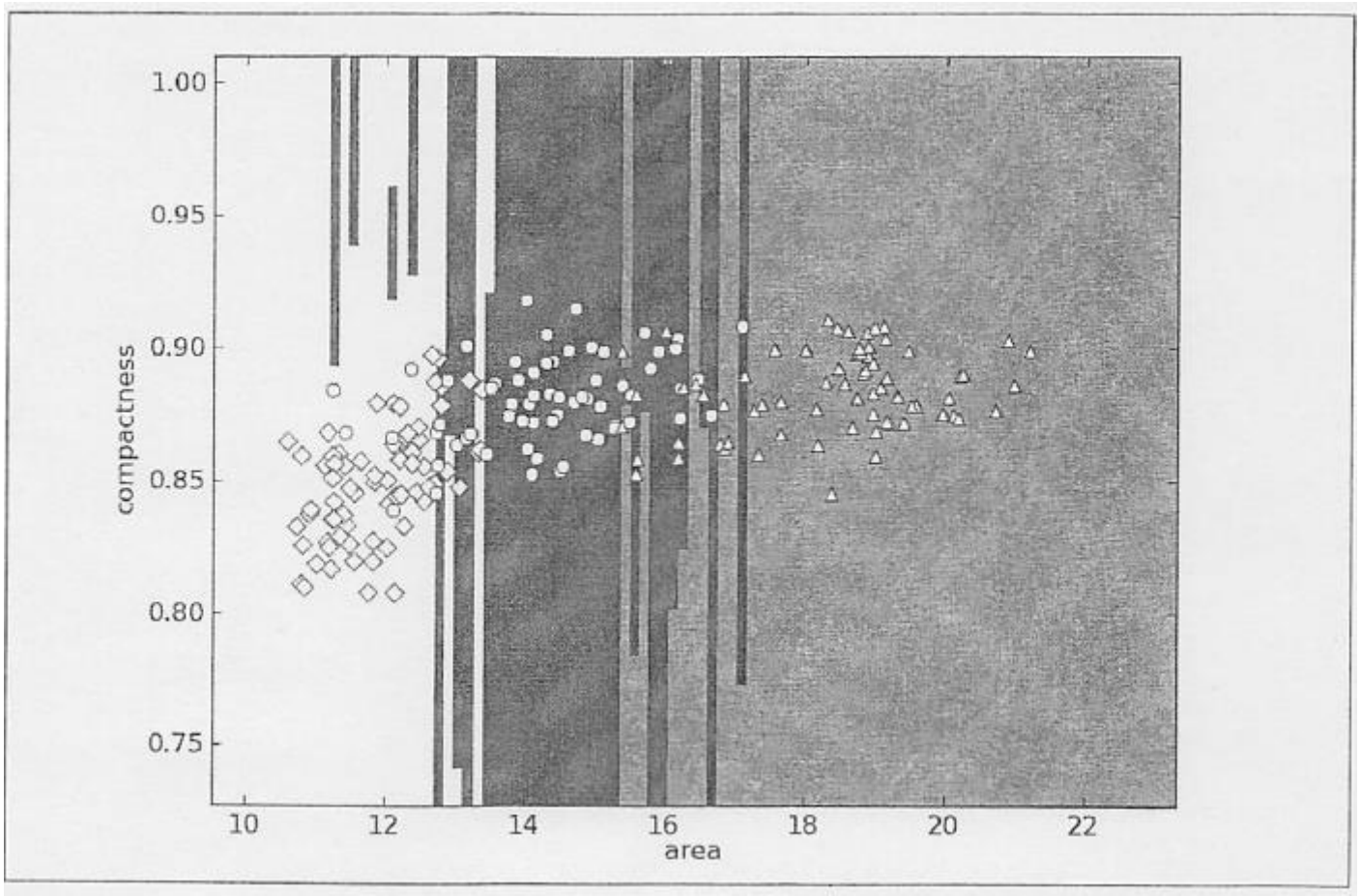
良い特徴量

- **重要なことには敏感に反応すること**
- **重要でないことには反応をしめさないこと**
- （この二つの条件を同時に満たす特徴量は求めることは難しい、理想に近づく特徴量を探索することが必要）

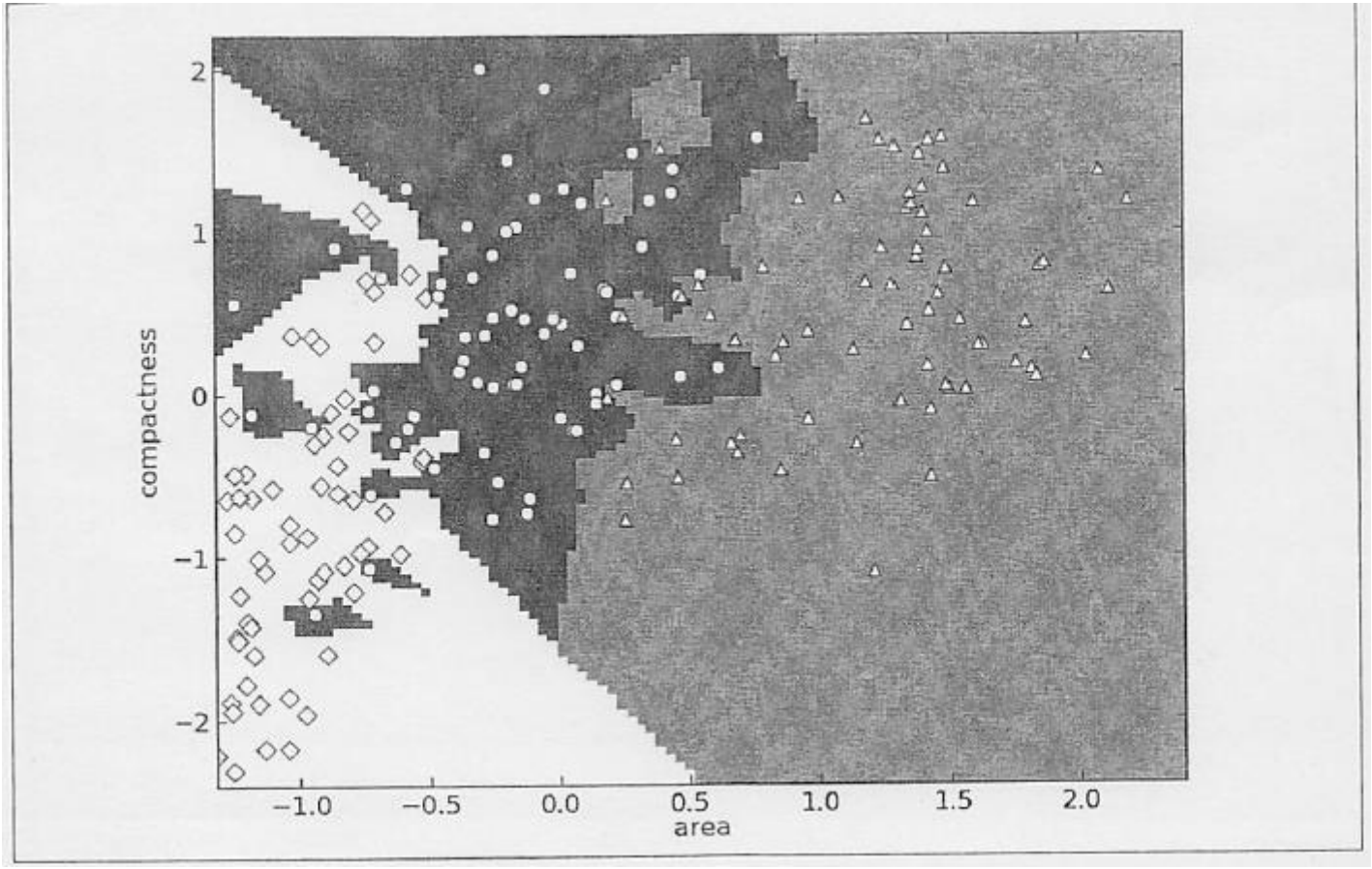
最近傍法

- 新しいデータが与えられた場合、そのデータに最も近い点をデータセットから探索し、その最近傍点のラベルを結果とするというものです。

面積と密集度の二つを軸にとってグラフ化

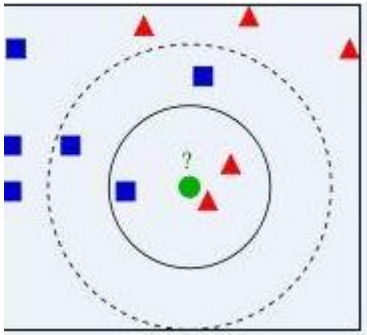


正規化したデータで最近傍法を用いると



k一近傍法

緑の円は赤い三角形のグループに属すべきか？或いは、青いの正方形のグループに属すべきか？

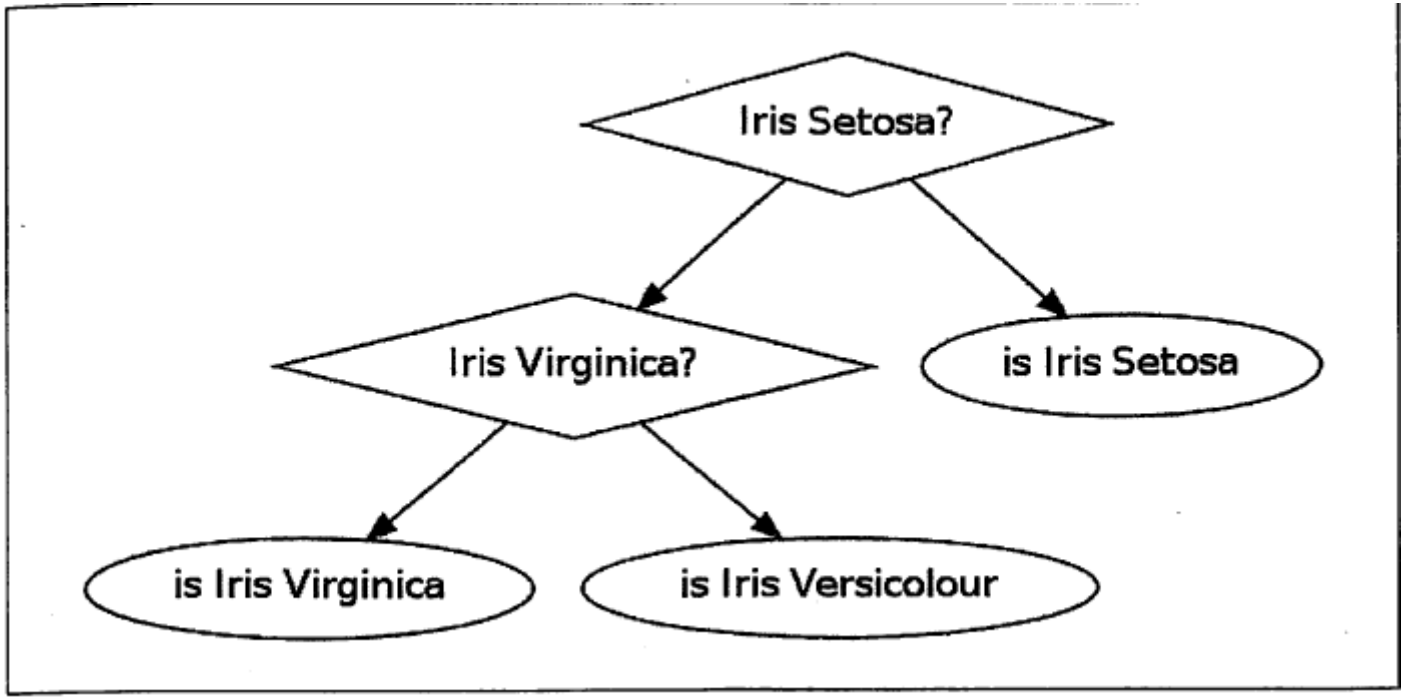


もしK＝3ならば、赤い三角形は全体の2/3を占めるので、緑の円は赤い三角形のグループに属すべきです。
もしK＝5ならば、青い正方形は全体の3/5を占めるので、緑の円は青い正方形のグループに属すべきです。

二項分類と多項分類

- 二項分類
- :分類したい対象にある条件によって分類を行う、あるクラスを属しているかどうかを判定します。（YES　OR　NO）
- 多項分類
- :クラスを複数が存在、分類したい対象にある条件によってどのクラスを属すべきかいうことを判定します

木構造の分類器


Iris Setosa?
Iris Virginica?
is Iris Setosa
is Iris Virginica
is Iris Versicolour

まとめ

- クラス分類を行うために、モデルを生成すること
- 訓練誤差を評価基準として使えません
- 訓練データで使用しなかったデータをテキストデータとして使うべき
- テストデータとしてデータを使い過ぎないために、交差検定を行うべき
- 特徴量を選択することも重要